# Improving Music Genre Classification
# EECE 5644: Machine Learning

Eric Doyle, doyle.e@husky.neu.edu
Timothy Rupprecht, rupprecht.t@husky.neu.edu

Instructor: Professor Jennifer Dy
TA: Fernando Quivira, Sadegh Salehi
Date Submitted: April 30, 2016

## 1 Abstract

Music genre classification is a growing testing ground for the field of machine learning. This project explores the benefits of using natural language processing on lyrics as well as signal processing of audio signals as means of genre classification for country, hip hop, and heavy metal genres. The project culminates in showing that using features from both datasets increases overall performance by 4.8 percentage points.

# 2 Introduction

Musical understanding occurs within a complex intersection of sound perception, social awareness, and historical perspective. Decisions that require musical understanding prove to be a difficult and rewarding task for artificial intelligence systems. Any music recommendation system requires the ability to create intelligent groupings of music, and classifying songs for the user. It is common to group songs into different musical genres, and research has led to increasingly accurate genre classifiers. Genre classifiers can perform well (most research reporting upwards of 85% accuracy) but work is always being done to improve these numbers.

Focusing on three music genres, we compared results of classification using text data against using signal data. Finally, we generated a system that combined both methods into a neural network to see if our accuracy increased when we combined both of these methods.

## 2.1 Objectives

From the very beginning of this project we grappled with working with something we loved and working with something that we personally thought was a more attainable achievement. We felt torn over whether we should work on natural language processing applied to twitter, or work on audio feature extraction to analyze music. However, after much consideration we decided that working with music would be much more preferable than attempting to do something for an easy grade. Over the course of researching genre classification two objectives emerged as possible goal posts to aim for.

1. Maximize accuracy for music genre identification.
2. Explore effectiveness of combining signal processing and natural language data as classifiers.

As one can see utilizing natural language processing still ended up being a major component of our project. There was very little on the internet showing that both lyric analysis as well as audio analysis had been combined in research attempts concerning genre classification. Depending on the results we would end up getting would dictate the project. If the results showed that the two did not work well together we intended to compare and contrast the strengths of the two different methods but we hoped to find a way of using both together to increase accuracy.

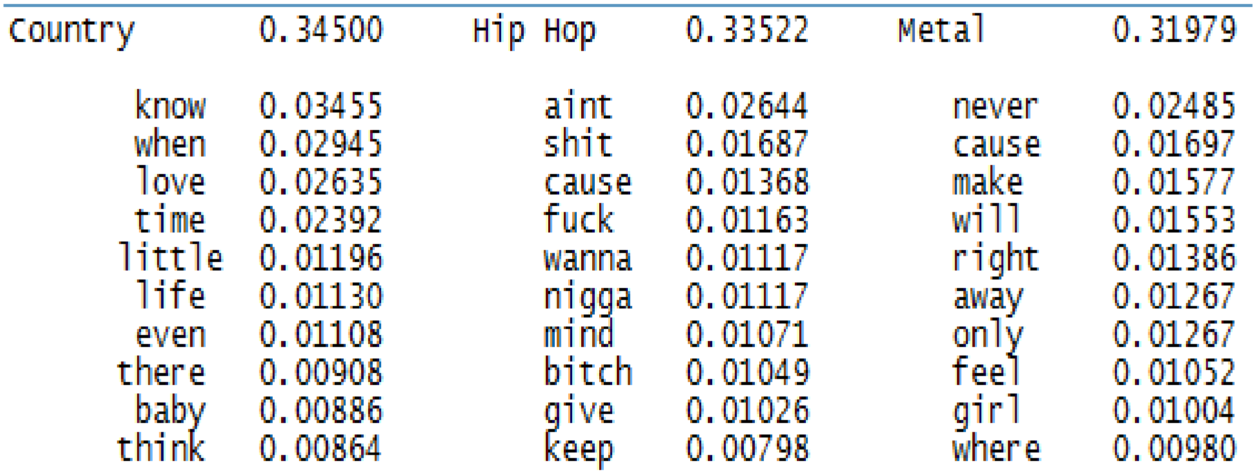
# 3 Multiple Approaches

## 3.1 Topic Model based Classifiers (on Lyrics)

In order to explore the effectiveness of how natural language processing could fair at genre classification we used a rather novel approach. Latent Dirichlet Allocation (LDA) was explored as an option for feature extraction of lyrics. It was known that LDA is good for when a smaller sample size is available for training. One matlab toolbox was discovered that utilized the Gibbs Sampling with Latent Dirichlet Allocation in order to classify news and research articles. The process is unsupervised and can be generalized to work with any sort of article one wishes to classify.

The novelty to our approach involved treating the various songs we had available from each genre as articles we wished to classify. Giving the algorithm the entire dataset we had it search out 3 topics. Our topics however corresponded to the three genres we hoped to classify: country, heavy metal, and hip hop. Just relying on intuition we knew that different genres would use different words with higher frequencies due to thematic influences of the genre.

Presented below are the results of the Gibbs Sampling which identifies key words in a topic. In our example, topics correspond to different genres.



```
Country        0.34500     Hip Hop     0.33522     Metal          0.31979

       know    0.03455        aint     0.02644        never       0.02485
       when    0.02945        shit     0.01687        cause       0.01697
       love    0.02635        cause    0.01368        make        0.01577
       time    0.02392        fuck     0.01163        will        0.01553
     little    0.01196        wanna    0.01117        right       0.01386
       life    0.01130        nigga    0.01117        away        0.01267
       even    0.01108        mind     0.01071        only        0.01267
      there    0.00908        bitch    0.01049        feel        0.01052
       baby    0.00886        give     0.01026        girl        0.01004
      think    0.00864        keep     0.00798        where       0.00980
```

From this method we saw pretty interesting results. For two genres (Hip Hop and Country) an accuracy of 82.5% was achieved. For three genres an accuracy of 64% was achieved. This result was slightly less than results seen in research papers using different analysis techniques but still performed better than the second method that was explored during the course of the project. The Gibbs Sampler assigns one of the three topics to each article based on which topic scored the highest for an individual sample. In our last natural language processing experiment we attempted to use these scores as inputs to a neural net. This method saw an accuracy of about 60% for three genres showing that using the Gibbs Sampler alone proved to be more reliable.

## 3.2 Waveform Neural Network

Based on the results of available research, we determined that neural networks are an excellent method for audio classification. In multiple papers, neural networks outperformed support vector machines (SVM) and classification and regression tree (CART), and unsupervised methods consistently underperform these supervised methods. SVMs and CART methods perform well with audio features, but for this project we wanted to use neural networks as a baseline method because it was the preferred method in a majority of research conclusions.

We constructed a two layer (one hidden layer) neural network to perform genre classification, with 17 audio features used as inputs and 3 music genre outputs. We leveraged the Neural Network Toolbox developed by Mathworks to generate the neural network in MATLAB. 70% of our dataset was used to train the network using error back propagation. Using the labels we gathered for all input samples, we first trained the network using backward error propagation to adjust the weights of our 40 neurons within the hidden layer. 15% of our dataset was used for validation, which checked that we were not overfitting our network to our dataset. Lastly, 15% of our data was used to test the network, and an accuracy was generated based on the outputs of the network and the true labels for those samples.

We obtained our audio sources from the MARSYAS (Music Analysis, Retrieval and Synthesis for Audio Signals), an open source music database collected by George Tzanetakis at the University of Victoria in Canada. This database provides 1000 songs from 10 different genres, but for our project only 3 genres and a total of 300 songs were utilized. Each audio sample from this database was 30 seconds long, and sampled at a rate of 22,050 kHz. The samples are taken from sources of varying audio fidelity (i.e. CD audio, streaming, radio). This database provides samples in the .au audio format, and conversion from .au to .wav had to be performed on every sample.

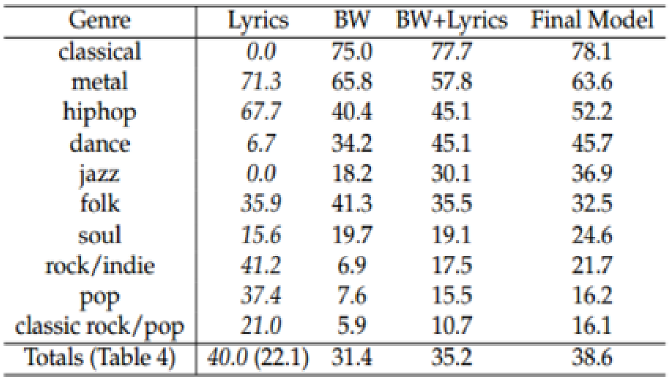
Using the MIRToolbox developed by the University of Jyvaskyla, we were able to extract audio features that distinguished musical audio most effectively. After experimentation, we concluded that 17 specific audio features were best for the inputs into our neural network. We began investigating different features based on musical intuition, and finalized our selection based on experimental results. The selected features were 13 Mel-frequency cepstral coefficients (MFCC), a value for loudness, the sum of the envelope, and two density values for calculated note onsets. The MIRToolbox provided functions to compute these values from the raw waveform of the sample.

MFCC is a popular algorithm for feature extraction on speech signals. This process converts the audio sample into evenly spaced mel-bands, which is filtered in a way similar to how humans perceive sound. In this manner, we thought it was appropriate to use this to decompose the spectrum of our audio sample to mimic the human auditory system response. Loudness was used as a measure to represent the overall volume of a sample. The energy used to compute loudness was generated from the total dB spectral energy represented using the bark scale. The bark scale are defined frequency bands that have been selected to also model human hearing. The total loudness was useful to differentiate quieter music genres from louder ones. Both of these features are derived from the spectrum of the waveform, which was calculated by taking the FFT of the waveform.

The time domain envelope of the waveform was calculated to derive our remaining features. We summed the magnitude of the resulting envelope to represent the overall magnitude within an audio sample. The MIRToolbox allowed us to classify note onsets, which are detected within the envelope of the sample. This information represents the moments when the envelope experiences a sharp variation, potentially indicating the beginning of a musical note. The onsets of notes is another feature that provides musical distinction between genres of music. Genres on average have a distinct density of notes being played within a time frame. Using the MIRToolbox mireventdensity(), we were given values corresponding to the density of note event onsets within a certain amount of time.

## 3.3 Combining Ideas

When looking at the work of others in the field of machine learning we saw very little indication that using both lyrical analysis as well as signal processing was being used together in a way that increased performance in genre classification. There is one paper from Dawen Liang, Haijie Gu, and Brendan OConnor called Music Genre Classification (2011) where they attempted something along those lines. In this experiment they used many more genres than our experiment intended to use, and they used different methodology in computing a final result for determining genre.



| Genre | Lyrics | BW | BW+Lyrics | Final Model |
|---|---|---|---|---|
| classical | *0.0* | 75.0 | *77.7* | 78.1 |
| metal | *71.3* | 65.8 | 57.8 | 63.6 |
| hiphop | *67.7* | 40.4 | 45.1 | 52.2 |
| dance | *6.7* | 34.2 | 45.1 | 45.7 |
| jazz | *0.0* | 18.2 | 30.1 | 36.9 |
| folk | *35.9* | 41.3 | 35.5 | 32.5 |
| soul | *15.6* | 19.7 | 19.1 | 24.6 |
| rock/indie | *41.2* | 6.9 | 17.5 | 21.7 |
| pop | *37.4* | 7.6 | 15.5 | 16.2 |
| classic rock/pop | *21.0* | 5.9 | 10.7 | 16.1 |
| Totals (Table 4) | *40.0* (22.1) | 31.4 | 35.2 | 38.6 |

This table shows accuracy as a percentage

From the above table you can see that in their experiments lyrics alone performed much better when the genres had lyrics to use as data than using audio signals alone. Combining the data yields about 4% better accuracy.

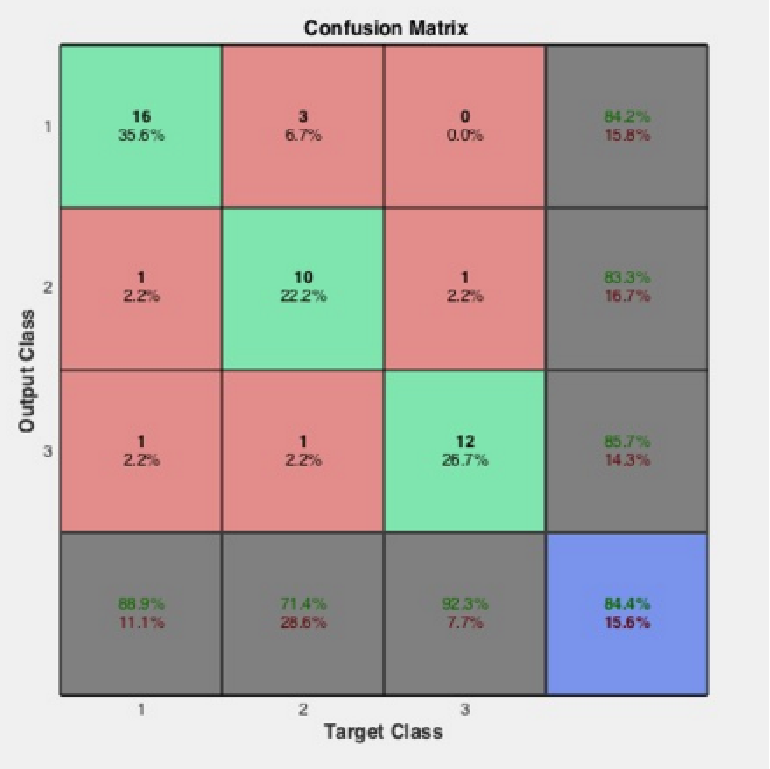
## 3.4 Results

Overall results were improved upon by using both audio signal features as well as the natural language features provided by the lyrics. The accuracy achieved is very impressive by the standards we have witnessed from contemporary research papers. Results for all of our experiments are available in the table seen below. It should be noted however that some research showed similar results with a much larger amount of genres being choosen from. More testing would need to be done on our own system would need to be done in order to truly compare our results to contemporary research.

| Methodology | Results |
|---|---|
| Gibbs sampling (Lyrics Only) | 64% |
| Topic model weights in neural network (Lyrics Only) | 60% |
| Audio Features in neural network (Audio Only) | 84% |
| Final System (Both Audio and Lyrics) | 88.8% |

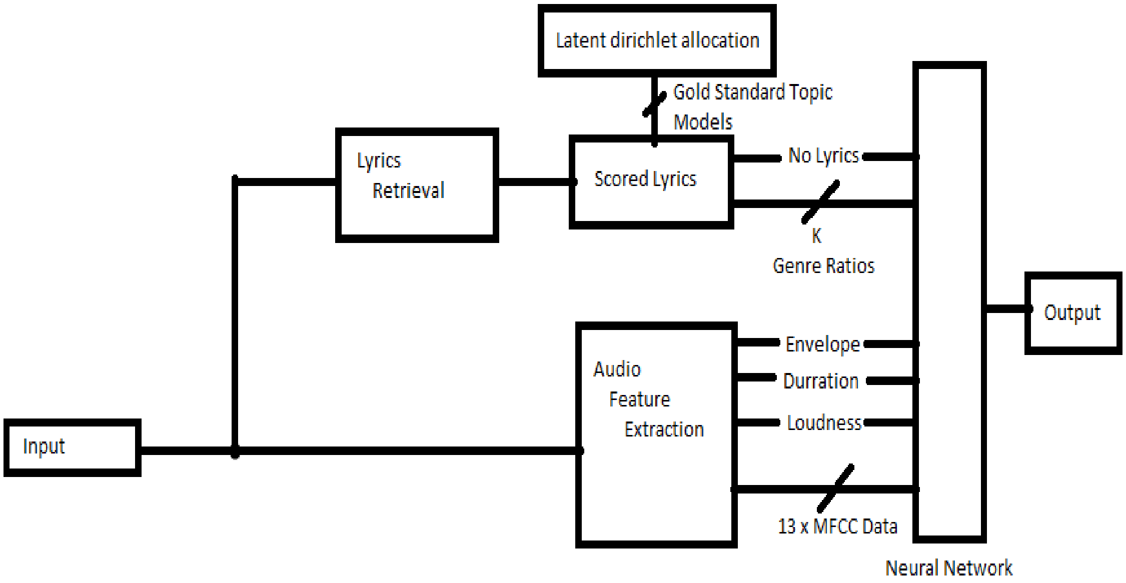
Table 1: Genre Classification on Hip Hop, Country, and Heavy Metal

Below you can see the confusion matrix for the combined approach.



# 4 System Overview

Combining the natural language processing and audio signal processing portions of the project proved to be relatively easy. Both portions of the system act in parallel unaware of each other. The outputs of each part are then fed into the same neural network that was used when attempting to classify genres using audio features alone.



Integrating the natural language processing proved to require the most work for this portion of the design of this stage of the project. In previous attempts different samples were used to train the Gibbs Sampler using LDA and the neural network used with audio feature extraction. In order to use the larger sample size required by the audio feature extraction a web crawler was utilized to retrieve lyrics for specific songs. These lyrics were then compared to what was referred to as a gold standard for each genre.

These gold standards were found using the Gibbs Sampling using LDA however we did not utilize the algorithm as a classifier. This makes this natural language processing slightly different from the stand alone lyrics analysis explored in previous stages of the project. The most drastic way it differs is how we use the LDA. As mentioned, it is used to create gold standards that represent the genres it operated on as a whole. Each genre was explored one at a time and 4 topics were derived per genre. It is these sets of topics that acted as our gold standard.

Using these gold standards each input to the algorithm gets compared to the topics that best represented each genre. Gibbs Sampling provides 10 words and their respective percentage of their inclusion in the samples as a whole. For new songs being analysed by the system each lyric would be compared to these topics for each genre and individual words from the song that matched with words from the topics in each genre would increase the score for that genre. All three scores were then sent as inputs to the neural network that compared all the features that our system extracted. This is why we believe the results are better for when you combine the two sets of data. Using the neural network the system could find training samples where the lyrics were implying one thing but there could still be a chance based on audio features that the song belonged to a different genre. This methodology brings genre classification using machine learning closer to the way humans perceive music.

# 5 Future Work

This project has succeeded as a proof of concept. We were able to increase the performance of our neural network by combining two different types of music related data. Audio and text are typically used independently for classification, and we have created evidence that they can improve a classifier when combined.

Improving our supervised classification method will help us achieve higher accuracy. Using a deep neural network might provide more accurate classification but will be cause to reconsider what audio features we use. The deep neural network will be more effective for high dimensional feature inputs. For now, the features we selected used in the two layer network perform well.

Leveraging text and audio data will create more effective learning methods to be used on multimedia data. For the example of audio, new musical content is almost entirely available through the internet on music streaming services, youtube, and social media. Traditional music genre classification is becoming outdated quickly. We need to focus on perfecting learning methods that combine the multimedia that is available on the internet, because this multimedia is reshaping and defining the musical genres themselves. Soundcloud is a music social platform and is an excellent example of where music is commonly described using hashtags or moods. New music will need to be classified using this textual data in addition to the audio waveform in order to create effective recommendation platforms for users. As music genres become more complex and more are numerous on the web, combining text processing and audio processing for feature extraction will become a requirement to achieve accuracy in classification problems.

# 6 Resources

[**1.** ] https://www.brandwatch.com/2013/08/social-media-the-music-industry/

[**2.** ] http://www.gizmag.com/automatic-music-genre-classification-system/38240/

[**3.** ] http://cs229.stanford.edu/proj2011/HaggbladeHongKao-MusicGenreClassification.pdf

[**4.** ] Matlab Topic Modeling Toolbox

[**5.** ] MIR Toolbox

[**6.** ] Marsyas Music Library